·论著·

基于麻雀搜索算法优化的 BP 神经网络模型对 2 型糖尿病肾病的预测研究

邹琼[1,2]，吴曦[1]，张杨[1]，万毅[3]，陈长生[1*]

1.710032 陕西省西安市，空军军医大学军事预防医学系军队卫生统计学教研室 特殊作业环境危害评估与防治教育部重点实验室
2.712046 陕西省咸阳市，陕西中医药大学公共卫生学院
3.710032 陕西省西安市，空军军医大学卫勤训练基地
*通信作者：陈长生，教授/博士生导师；E-mail：chencs@fmmu.edu.cn

【摘要】 **背景** 糖尿病肾病（DN）是糖尿病常见的微血管并发症之一，发病率高，危害性大。早期发现 DN 对预防相关疾病非常重要。目前大多研究基于传统的统计预测方法，数据需满足其所要求的前提假设条件。近年来已无法很好满足其在 DN 预测领域的需求，有必要尝试开展机器学习等新方法在 DN 预测领域的应用。**目的** 利用 LASSO 回归和麻雀搜索算法（SSA）优化的 BP 神经网络（SSA-BP 神经网络）构建 DN 预测模型。**方法** 本研究时间为 2023 年 4—8 月，数据来源于公开的伊朗 133 名糖尿病患者的并发症数据。采用 SPSS 26.0 软件进行单因素分析，采用 LASSO 回归筛选变量。以是否患 DN 为因变量，分别用 8∶2 和 7∶3 的比例划分训练集和测试集，使用 SSA-BP 神经网络进行建模与分析，并与经典的机器学习模型对比预测性能以分析较优的 DN 模型。基于准确率、精确率、灵敏度、特异度、F1-score 和 AUC 指标进行模型评价。**结果** 剔除 9 例 1 型糖尿病患者，本研究纳入的有效样本量为 124 例 2 型糖尿病（T2DM）患者，其中 73 例（58.9%）被诊断为 DN 患者。单因素分析显示年龄、BMI、糖尿病持续时间、空腹血糖（FBS）、糖化血红蛋白（HbA_{1c}）、低密度脂蛋白（LDL）、高密度脂蛋白（HDL）、三酰甘油（TG）、收缩压（SBP）和舒张压（DBP）的 T2DM 患者 DN 危险因素（$P<0.05$）。训练集∶测试集 =8∶2 时，训练集（n=100）中有 59 例 DN 患者，测试集（n=24）含有 14 例 DN 患者。LASSO 回归筛选出年龄、糖尿病持续时间、HbA_{1c}、LDL 和 SBP 共 5 个影响因素。Logistic 回归（LR）、K 近邻（KNN）、支持向量机（SVM）、BP 神经网络、SSA-BP 神经网络模型在测试集的准确率分别为 83.33%、79.17%、79.17%、87.50%、95.83%。F1-score 分别为 0.846 2、0.800 0、0.800 0、0.888 9、0.960 0。训练集∶测试集 =7∶3 时，训练集（n=88）中有 52 例 DN 患者，测试集（n=36）含有 21 例 DN 患者。LASSO 回归筛选出年龄、BMI、糖尿病持续时间、LDL、HDL、SBP 和 DBP 这 7 个影响因素。LR、KNN、SVM、BP 神经网络、SSA-BP 神经网络模型在测试集的准确率分别为 86.11%、86.11%、86.11%、72.22%、91.67%。F1-score 分别为 0.871 8、0.871 8、0.864 9、0.705 9、0.909 1。**结论** LR、KNN 和 SVM 模型在训练集∶测试集 =7∶3 时性能较好，BP 神经网络和 SSA-BP 神经网络模型在训练集∶测试集 =8∶2 时性能较好。相较于 BP 神经网络模型和传统机器学习模型，SSA-BP 神经网络模型的预测性能更佳，可及时准确识别 T2DM DN 患者，实现 DN 的早发现和早治疗，从而预防并减缓对其身体带来的危害。

【关键词】 糖尿病，2 型；糖尿病肾病；神经网络，计算机；预测模型
【中图分类号】 R 587.1 【文献标识码】 A DOI：10.12114/j.issn.1007-9572.2023.0360

Prediction of Type 2 Diabetic Nephropathy Based on BP Neural Network Optimized by Sparrow Search Algorithm

ZOU Qiong[1,2]，*WU Xi*[1]，*ZHANG Yang*[1]，*WAN Yi*[3]，*CHEN Changsheng*[1*]
1.Department of Military Health Statistics，School of Military Preventive Medicine，Air Force Medical University/Ministry of Education Key Lab of Hazard Assessment and Control in Special Operational Environment，Xi´an 710032，China
2.College of Health Public，Shaanxi University of Chinese Medicine，Xianyang 712046，China

基金项目：国家自然科学基金资助项目（82073663）
引用本文：邹琼，吴曦，张杨，等.基于麻雀搜索算法优化的 BP 神经网络模型对 2 型糖尿病肾病的预测研究［J］.中国全科医学，2023.［Epub ahead of print］.DOI：10.12114/j.issn.1007-9572.2023.0360.［www.chinagp.net］
ZOU Q，WU X，ZHANG Y，et al. Prediction of type 2 diabetic nephropathy based on BP neural network optimized by sparrow search algorithm［J］. Chinese General Practice，2023.［Epub ahead of print］.
本文数字出版日期：

3.Department of Health Services, Air Force Medical University, Xi´an 710032, China
*Corresponding author: CHEN Changsheng, Professor/Doctoral supervisor; E-mail: chencs@fmmu.edu.cn

【Abstract】 **Background** Diabetic nephropathy (DN) is one of the most common microvascular complications of diabetes, which is highly prevalent and harmful. Early detection of DN is an important task in preventing related diseases. Currently, most of the researches are based on traditional statistical prediction methods, and data need to meet the prerequisites it requires. It is necessary to try to apply new methods such as machine learning in the area of DN prediction for its failing to meet the needs in the field of DN prediction in recent years. **Objective** To construct DN prediction model using the LASSO regression and BP neural network optimized by sparrow search algorithm (SSA-BP). **Methods** This study was conducted from April 2023 to August 2023, and the data was obtained from publicly available data on complications of 133 patients with diabetes mellitus in Iran. Univariate analysis was conducted using SPSS 26.0 software, and variables were screened using LASSO regression. Using the presence of DN as the dependent variable, the training and testing sets were divided into 8 : 2 and 7 : 3 ratios, respectively. The SSA-BP neural network was used for modeling and analysis, and the prediction performance was compared with classical machine learning models to analyze the better DN model. Model evaluation was performed based on accuracy, precision, sensitivity, specificity, F1-score and AUC indicators. **Results** Excluding 9 patients with type 1 diabetes, the effective sample size included in this study was 124 patients with type 2 diabetes mellitus (T2DM), of which 73 (58.9%) were diagnosed with DN. Univariate analysis of risk factors for type 2 DN showed statistically significant for age, BMI, duration of diabetes, fasting blood glucose (FBG), glycosylated hemoglobin (HbA_{1c}), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triacylglycerol (TG), systolic blood pressure (SBP) and diastolic blood pressure (DBP) ($P<0.05$). When the ratio of the training set to the test set was 8 : 2, there were 59 DN patients in the training set ($n=100$) and 14 DN patients in the test set ($n=24$). Five influencing factors of age, diabetes duration, HbA_{1c}, LDL, and SBP were obtained by LASSO regression screening. The accuracy rates of Logistic regression (LR), K-nearest neighbor (KNN), support vector machine (SVM) and SSA-BP models in the test set were 83.33%, 79.17%, 79.17%, 87.50%, and 95.83%, with F1-score as 0.846 2, 0.800 0, 0.800 0, 0.888 9, and 0.960 0, respectively. When the ratio of the training set to the test set was 7 : 3, there were 52 DN patients in the training set ($n=88$) and 21 DN patients in the test set ($n=36$). Seven influencing factors obtained by LASSO regression screening included age, BMI, diabetes duration, LDL, HDL, SBP, and DBP. The accuracy rates of LR, KNN, SVM, BP, and SSA-BP models in the test set were 86.11%, 86.11%, 86.11%, 72.22%, and 91.67%, with F1-score as 0.871 8, 0.871 8, 0.864 9, 0.705 9, and 0.909 1, respectively. **Conclusion** LR, KNN, and SVM perform better when the training set to the test set is 7 : 3, while BP and SSA-BP perform better when the training set to the test set is 8 : 2. Compared with the BP neural network and traditional machine learning models, SSA-BP model has the best prediction performance and can timely and accurately identify type 2 DN patients, realize early detection and treatment of DN, thus preventing and mitigating the harm to their bodies.

【Key words】 Diabetes mellitus, type 2; Diabetic nephropathies; Neural networks, computer; prediction model

糖尿病是最常见的人类疾病之一，已成为世界范围内重要的公共卫生问题[1]。糖尿病肾病（DN）是2型糖尿病（T2DM）常见的慢性微血管并发症，也是世界范围内终末期肾病（ESRD）的主要原因。印度、中国及其他发展中国家受糖尿病影响的人数正在迅速增长，给患者和卫生保健系统造成了世界性的负担[2]。因此，实现 DN 的早期诊断和治疗，有助于预防或延缓其发生、发展，从而提高患者的预期寿命[3]。

为了更好地控制疾病的进程，诊断出更易患 DN 的患者至关重要[3]。近年来，随着数据挖掘的发展，机器学习在糖尿病研究中发挥着越来越重要的作用[4]。其中 K 近邻（KNN）、支持向量机（SVM）和反向传播神经网络（BPNN）模型是常见的数据挖掘模型。与 SVM 等传统的机器学习算法相比，BP 神经网络具有良好的非线性映射能力、自适应性、容错性等优点[5]，但在实际应用中也存在一定缺陷，如易陷入局部极小值、结果存在随机性、网络收敛速度慢等[6]。因此，有必要改进标准的 BP 神经网络算法。麻雀搜索算法（SSA）是 XUE 等[7]受麻雀觅食和反捕食行为为启发而提出的一种仿生智能优化算法，因其具有良好的灵活性和全局寻优能力，研究者们将其与 BP 神经网络相结合以弥补其缺点，但目前多应用于电力工业、自动化技术等领域[8-9]。因此本研究将探索 SSA 优化的 BP（SSA-BP）神经网络应用于 DN 的诊断预测中，以期提升模型预测的准确率，或可为 DN 的早期筛查和诊断治疗提供理论依据 / 临床参考。

1 对象与方法

1.1 研究对象

数据来源于 KHODADADI 等[10]公开的伊朗 133 名



糖尿病患者的并发症数据（https：//data.mendeley.com/datasets/k62fdsnwkg/1）。数据集由133例糖尿病患者（1型和2型）的24项信息组成：性别、年龄、BMI、糖尿病类型、糖尿病持续时间、空腹血糖（FBG）、糖化血红蛋白（HbA_{1c}）、低密度脂蛋白（LDL）、高密度脂蛋白（HDL）、三酰甘油（TG）、治疗类型、他汀类药物类型、他汀类药物剂量、神经病变、肾病、视网膜病变、周围血管疾病、心血管疾病、足部溃疡、黎明效应、收缩压（SBP），舒张压（DBP）、累积阿托伐他汀当量（实际低密度脂蛋白胆固醇。依据既往文献[10-12]，提取了13个可能与DN患者相关的风险因素，变量赋值见表1。

表1　变量赋值说明

Table 1　The description of variable assignment

编号	变量名	赋值情况及值范围
1	肾病	否=0（对照），是=1
2	性别	女=0（对照），男=1
3	年龄（岁）	<40=1（对照），40~<60=2，≥60=3
4	BMI（kg/m^2）	<18.5=1（对照），18.5~<24.0=2，24.0~<28.0=3，≥28.0=4
5	糖尿病持续时间（年）	<10=0（对照），≥10=1
6	FBS（mg/dL）	实测值：80~510
7	HbA_{1c}（mg/dL）	实测值：6.5~13.3
8	LDL（mg/dL）	实测值：36~267
9	HDL（mg/dL）	实测值：20~62
10	TG（mg/dL）	实测值：74~756
11	治疗类型	口服剂=1（对照），胰岛素=2，二者=3
12	他汀类药物类型	无他汀类药物=1（对照），阿托伐他汀=2，瑞舒伐他汀=3
13	SBP（mmHg）	实测值：105~180
14	DBP（mmHg）	实测值：60~120

注：FBG=空腹血糖，HbA_{1c}=糖化血红蛋白，LDL=低密度脂蛋白，HDL=高密度脂蛋白，TG=三酰甘油，SBP=收缩压，DBP=舒张压；1 mmHg=0.133 kPa。

1.2　数据处理

剔除9例1型糖尿病患者，本研究纳入的有效样本量为124例T2DM患者，其中73例患有DN。对年龄、糖尿病持续时间和BMI连续变量离散化并编码。以是否患DN为因变量，分别用8∶2和7∶3的比例划分训练集和测试集。

1.3　研究方法

将单因素分析（表2）中$P<0.05$的变量纳入LASSO回归中进一步筛选并确定最终纳入模型的变量，在训练集上分别使用Logistic回归（LR）、KNN、SVM、BP神经网络和SSA-BP神经网络建立DN预测模型，并在测试集上进行验证。

1.4　方法学介绍

1.4.1　LASSO回归是TIBSHIRANI[13]提出的一种著名的稀疏回归方法。作为一种变量选择方法，LASSO回归需要一个惩罚项来约束系数的大小，并最终将结构风险降至最低，防止“过拟合”发生[14]。筛选的方法主要包括lambda.min和lambda.1se。因λ到达一定值之后，继续增加自变量个数并不能很显著地提高模型性能，而lambda.1se（距离均方误差一个标准误时的λ值）可给出一个具备优良性能且自变量个数最少的模型[15]。

1.4.2　KNN算法是一种监督机器学习算法，可用于解决回归和分类问题[16]。KNN分类是最基本、最简单的分类方法之一，在对数据分布知之甚少或一无所知的情况下，该方法是分类研究的首选方法之一。其不需要考虑模型构建的细节，且模型中唯一可调整的参数是K[17]。其易于理解和实现，但主要缺点是随着使用中数据的大小增长，速度会明显变慢[16]。

1.4.3　CORTES等[18]于1995年提出了SVM模型。SVM的常见的核函数种类有：线性、多项式、高斯和Sigmoid核函数。优点是泛化错误低，可获得准确和稳健的结果，适用于非结构化和半结构化数据集（如图像和文本）。缺点是当用于大型学习任务时，对内存和时间要求较高[19]。此外，对参数调节和核函数的选择敏感，变量的权重在最终模型中难以解释[20]。

1.4.4　BP神经网络是根据误差反向传播算法训练的多层前馈网络，是应用较广泛的神经网络模型之一。SSA在一定程度上改进了对优化搜索空间的探索和利用，有效地避免了局部最优问题[7]。在觅食过程中，麻雀被分为发现者、加入者和预警者[7]。假设d维空间中有N只麻雀，每只麻雀的位置为X=［x_1，χ^2，…，x_D］，适应度值f=f（x_1，χ^2，…，x_D）。该算法主要分为3部分，通过3个公式来进行更新[22]。首先，发现者的位置更新如公式（1）：

$$X_{i,j}^{t+1}=\begin{cases}X_{i,j}^{t}\cdot exp\left(\frac{-i}{\alpha\cdot iter_{max}}\right) & if\ R_2<ST\\ X_{i,j}^{t}+Q.L & if\ R_2\geqslant ST\end{cases}\qquad 公式（1）$$

式中t表示当前迭代次数，j=1，2，…，d，$X_{i,j}^{t}$表示迭代t时第i个麻雀的第j维的值。$iter_{max}$是最大迭代次数（常数），α是一个随机数（$\alpha\in(0,1]$）。R_2（$R_2\in[0,1]$）和ST（$ST\in[0.5,1.0]$）分别表示预警值和安全值。Q是服从正态分布的随机数。L表示一个1×d维的矩阵，其内部每个元素都为1。

其次，加入者的位置更新如公式（2）：

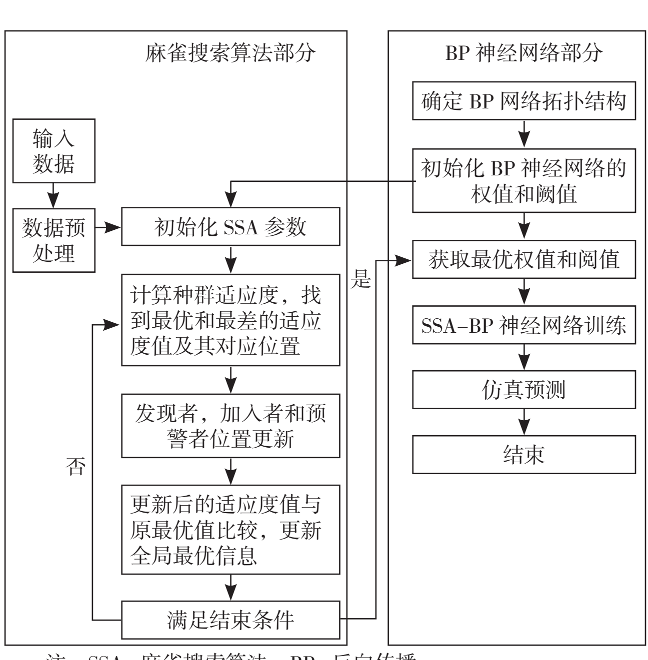
$$X_{i,j}^{t-1}=\begin{cases}Q\cdot exp\left(\frac{X_{worst}^{t}-X_{i,j}^{t}}{i^2}\right) & if\ i>\frac{n}{2}\\ X_p^{t+1}+|X_{i,j}^{t}-X_p^{t+1}|.A^{+}.L & otherwise\end{cases}\qquad 公式（2）$$

X_P是发现者占据的最优位置。X_{worst}表示当前全局最差位置。$A^+=A^T(AA^T)^{-1}$，A表示1×d的矩阵，其中每个元素随机分配1或-1，A的转置是A^T。当i＞n/2时，

表明第 i 个适应度值较差的加入者最有可能处于饥饿状态。预警者一般占总种群的 10%-20%，这些麻雀的初始位置是在种群中随机生成的，其位置更新如公式(3):

$$X_{i,j}^{t-1}=\begin{cases}X_{beset}^{t}+\beta\,.|X_{i,j}^{t}-X_{beset}^{t}| & if\ f_i>f_g\\ X_{i,j}^{t}+K.\left(\dfrac{|X_{i,j}^{t}-X_{worst}^{t}|}{(fi-fw)+\varepsilon}\right) & if\ fi=f_g\end{cases}\quad \text{公式（3）}$$

式（3）中 X_{best} 为当前全局最优位置，代表种群中心的位置，并且在它周围是安全的。β 是服从均值为 0 方差为 1 的正态分布的随机数，作为步长控制参数。K（$K\in[-1,1]$）是一个随机数，表示麻雀移动的方向，也是一个步长控制参数。f_i 是当前麻雀的适应度值，f_g 和 f_w 分别是当前全局最佳和最差适应度值，ε 是常数[21]。图 1 是算法的流程图。



注：SSA= 麻雀搜索算法，BP= 反向传播。

图 1 SSA-BP 神经网络流程图

Figure 1 Flow chart of SSA-BP neural network

1.5 统计分析与软件

采用 SPSS 26.0 软件进行统计学分析，以 $P<0.05$ 为差异有统计学意义。计数资料采用［例（%）］描述，两组比较采用 χ^2 检验或 Fisher's 确切概率法。符合正态分布的计量资料以（$\bar{x}\pm s$）表示，两组间比较采用两独立样本 t 检验。非正态分布的计量资料用 M（QR）表示，两组间比较采用 Mann-Whitney 检验。采用 R 4.2.2 软件中的 glmnet、kknn、e1071 程序包在训练集上建立 LASSO 回归、KNN 和 SVM 模型。采用 caret 程序包的 dummyVars 函数对多分类变量进行哑变量处理。采用 MATLAB 2022a 软件构建 BP 神经网络和麻雀搜索 SSA-BP 的神经网络模型。最后在测试集上评价性能，由混淆矩阵计算出的准确度、精确度、灵敏度和特异度来判断各模型的优劣。

2 结果

2.1 一般资料

124 例研究对象中 73 例（58.9%）被诊断为 DN。

2.2 T2DM 患者 DN 危险因素的单因素分析

无 DN 和患 DN 患者的性别、治疗类型、他汀类药物类型比较，差异无统计学意义（$P>0.05$）；无 DN 和患 DN 患者的年龄、BMI、糖尿病持续时间、FBS、HbA_{1c}、LDL、HDL、TG、SBP、DBP 比较，差异有统计学意义（$P<0.05$），见表 2。

2.3 LASSO 回归变量筛选

基于训练集，以是否发生 DN 为因变量，以单因素分析中有统计学意义的 10 个变量为自变量进行 LASSO 回归分析。多分类变量在纳入模型前先进行哑变量化(10 个自变量变为 13 个候选变量)。选择 10 倍交叉验证下 lambda.1se(lambda.1se=0.068 191 87)为模型最优值(图 2)，训练集：测试集 =8：2 时结果显示，年龄、糖尿病持续时间、HbA_{1c}、LDL 和 SBP 是与 DN 发生相关的 5 个变量，训练集：测试集 =7∶3 时结果显示，年龄、BMI、糖尿病持续时间、LDL、HDL、SBP 和 DBP 是与 DN 发生相关的 7 个变量。

2.4 LR 模型的建立

以是否发生 DN 为因变量（赋值：否 =0，是 =1），以 LASSO 回归筛选出的变量为自变量进行 LR 分析。多分类变量进行哑变量处理，因某些分类算法（如 SVM、LR 和神经网络）在未缩放的数据上表现不佳[22]，所以计量资料采用标准化公式归一化处理成(0，1)区间的变量，进而建立 LR 模型。训练集：测试集 =8：2 时，结果显示糖尿病持续时间（OR=6.615，95%CI=1.263~42.533）、LDL（OR=3.647，95%CI=1.493~10.511）、SBP（OR=4.884，95%CI=1.863~17.332）是 DN 的危险因素（$P<0.05$）。LR 模型表达式为 Logit（P）=1.861+1.889×糖尿病持续时间 +1.294×LDL+1.586×SBP（R^2=0.767）。训练集：测试集 =7∶3 时，糖尿病持续时间（OR=6.786，95%CI=1.154~54.104）、LDL（OR=5.834，95%CI=2.128~21.033）是 DN 的危险因素（$P<0.05$），表达式为 Logit（P）=-16.041+1.915×糖尿病持续时间 +1.764×LDL（R^2=0.739）。

2.5 KNN 模型的建立

以是否发生 DN 为因变量（赋值：否 =0，是 =1）在训练集上建立 KNN 模型。利用 Caret 包中 train（）函数的网格搜索法寻找 K 的最佳参数，K 的初始取值范围为［2，15］。分别在训练集：测试集 =8：2 和 7∶3 时，十折交叉验证正确率最高时得到的最优 K 值分别为 14 和 9。

表 2　2 型糖尿病肾病患者相关危险因素的单因素分析

Table 2　Univariate analysis of risk factors associated with type 2 diabetic nephropathy

变量	无 DN（n=51）	患 DN（n=73）	检验统计量值	P 值
性别［例（%）］			1.759[a]	0.185
女	34（66.7）	40（54.8）		
男	17（33.3）	33（45.2）		
年龄［例（%）］			19.229[a]	<0.001
<40 岁	5（9.8）	4（5.5）		
40~<60 岁	37（72.5）	28（38.4）		
≥ 60 岁	9（17.6）	41（56.2）		
BMI［例（%）］			13.100[a]	0.002
<18.5 kg/m^2	2（3.9）	0		
18.5 ~<24.0 kg/m^2	10（19.6）	2（2.7）		
24.0~<28.0 kg/m^2	9（17.6）	12（16.4）		
≥ 28.0 kg/m^2	30（58.8）	59（80.8）		
糖尿病持续时间［例（%）］			27.358[a]	<0.001
<10 年	39（76.5）	21（28.8）		
≥ 10 年	12（23.5）	52（71.2）		
FBG（$\bar{x} \pm s$，mg/dL）	181.33 ± 65.97	229.03 ± 54.84	−4.381[b]	<0.001
HbA_{1c}［M（QR），%］	8.10（1.60）	10.80（0.95）	−5.773[c]	<0.001
LDL（$\bar{x} \pm s$，mg/dL）	109.12 ± 35.17	152.68 ± 42.672	−6.003[b]	<0.001
HDL（$\bar{x} \pm s$，mg/dL）	38.55 ± 8.43	35.74 ± 5.836	2.193[b]	0.030
TG（$\bar{x} \pm s$，mg/dL）	181.96 ± 84.95	242.04 ± 102.793	−3.433[b]	0.001
DM_treat［例（%）］			4.281[a]	0.113
口服剂	35（68.6）	38（52.1）		
胰岛素	4（7.8）	14（19.2）		
二者	12（23.5）	21（28.8）		
Statin［例（%）］			0.814[a]	0.778
无他汀类药物	16（31.4）	19（26.0）		
阿托伐他汀	34（66.7）	53（72.6）		
瑞舒伐他汀	1（2.0）	1（1.4）		
SBP（$\bar{x} \pm s$，mmHg）	130 ± 15	155 ± 14	−9.524[b]	<0.001
DBP（$\bar{x} \pm s$，mmHg）	81 ± 9	98 ± 12	−8.499[b]	<0.001

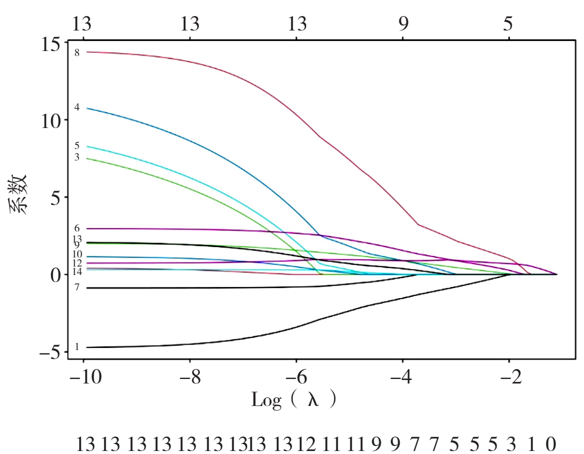
注：DN= 糖尿病肾病；[a] 表示 χ^2 值，[b] 表示 t 值，[c] 表示 Z 值。

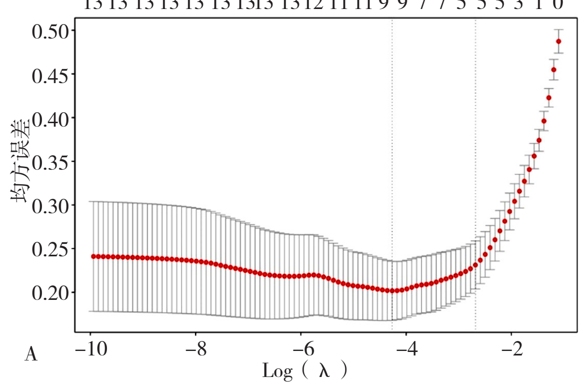
2.6　SVM 模型的建立

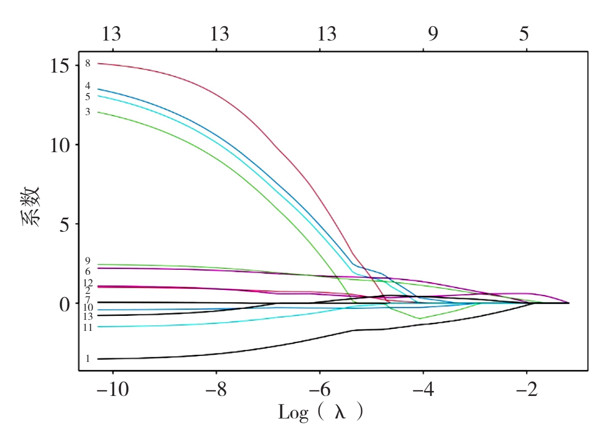
以是否患 DN 为因变量（赋值：否 =0，是 =1），LASSO 回归筛选的变量为自变量（如表 1）建立径向基核函数支持向量机模型（kernel=" radial"）。利用 R 软件中的 tune.svm（）函数的网格搜索法来寻找最优参数，C 与 γ 的初始取值范围分别为［0.001，0.01，0.1，1，5，10，100，1 000］和［0.1，0.5，1，2，3，4］。在训练集：测试集 =8∶2 和 7∶3 时，十折交叉验证错误率最低时的选出的最佳参数分别为 C=10、γ=0.1 和 C=1、γ=0.1。

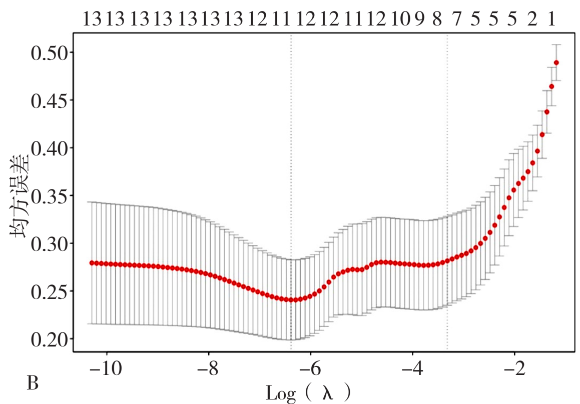
2.7　BP 神经网络模型的建立

考虑到训练时间和模型复杂度，本研究建立 3 层 BP 神经网络模型。将样本值进行归一化处理，这有助于提高网络的训练速度。在训练集：测试集 =8：2 时，









注：A 训练集：测试集 =8 ：2 时，13 个变量的系数曲线和 10 倍交叉验证的 LASSO 回归选择最佳的变量；B 训练集：测试集 =7 ：3 时，13 个变量的系数曲线和 10 倍交叉验证的 LASSO 回归选择最佳的变量

图 2　LASSO 回归变量筛选

Figure 2　LASSO regression screening for variables

输入层节点数（n）为5，输出层节点数（m）为2。基于常用的经验公式[23]：［$h=\sqrt{n+m}+a$，$a\in$（1，10）］，根号7为2.6，再加上 a，则隐藏层节点范围［3.6，12.6］，则取[3，12]。同理在训练集：测试集 =7：3 时，n 为7，m 为2，隐藏层节点范围则为[4，13]。经多次试验，在训练集：测试集 =8：2 和 7：3 时，最佳隐藏层节点数分别为 8 和 12 时拟合效果最好，因此网络拓扑结构分别设为 5-8-2 和 7-12-2。隐藏层及输出层的激励函数采用双曲正切 S 型函数及线性求和函数：$tansig(n)=2/(1+e^{-2n})-1$；$purelin(n)=n$，训练次数 1 000 次，网络训练速率为 0.01，训练目标最小误差为 0.000 1，Levenberg-Marquardt 法为训练算法，用梯度下降法更新权重。

2.8　SSA-BP 神经网络模型的建立

参数初始化：SSA 的进化代数为 50，种群规模为 30，安全值 ST 为 0.6；发现者比例 PD 为 0.7，意识到有危险的麻雀的比重 SD 为 0.2；适应度函数设计为训练集与测试集整体准确率的平均值，适应度函数值越大，表明模型训练越准确，随后计算个体适应度；更新发现者、加入者和预警者的位置；查看位置更新之后的个体适应度，并与当前最优适应度值进行比较，达到最终止条件则选择全局最优解；否则，再次进行迭代；将模型输出的最优解作为神经网络的权值和阈值，代入 BP 神经网络进行训练，利用误差反向传播调节参数，当达到最大迭代次数（1 000）或目标误差（0.000 1）的时，训练停止；将 SSA 优化后的 BP 神经网络模型用于预测是否患 DN。用均方误差（MSE）表示模型性能，训练集：测试集 =8：2 和 7：3 时，最终模型分别在迭代 20 和 5 次处达到最优。模型进化 / 适应度曲线表明模型在不断优化，最终达到最佳的适应度值（图 3）。

2.9　模型比较

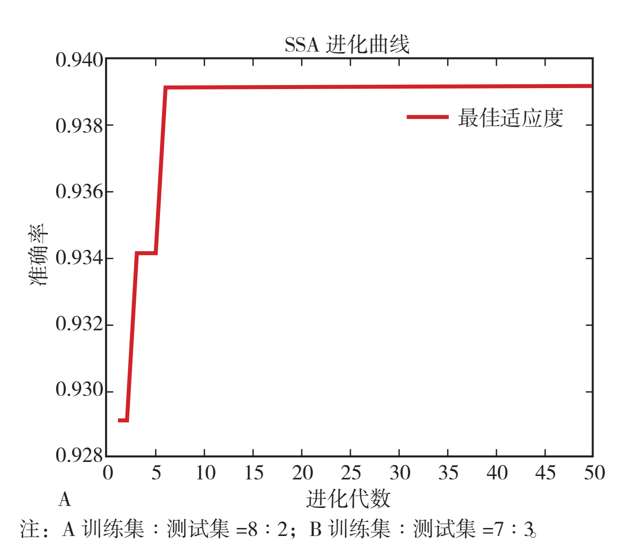
基于测试集验证上述模型的预测性能，各模型训练集与测试集的结果见表 3。在训练集：测试集 =8：2 时，KNN 模型和 SVM 模型达到了同样的性能，其在训练集上优于 LR 模型，但在测试集上却不如 LR 模型。BP 模型在测试集上的准确率，灵敏度，F1-score 和 AUC 优于 LR 模型，KNN 模型和 SVM 模型，整体上 SSA-BP 模型在训练集和测试集上的性能优于 BP、LR、KNN、SVM 模型。

在训练集：测试集 =7：3 时，LR 模型和 KNN 模型在测试集上结果相同，但在训练集上 KNN 模型性能优于 LR 模型。LR、KNN、SVM 模型在测试集上具有相同的准确率，但 SVM 模型的精确率，特异度和 AUC 高于 LR 模型和 KNN 模型。不管在训练集还是测试集上，LR、KNN、SVM 模型的效能优于 BP 网络神经网络。SSA-BP 网络神经模型提高了 BP 神经网络模型的性能，使得测试集上 BP 神经网络模型的准确率从 72.22% 提升到了 91.67%。

对比 2 个划分比例下的模型性能，发现 LR、KNN、SVM 模型在训练集：测试集 =7：3 时预测性能较高，而 BP 和 SSA-BP 则在训练集：测试集 =8：2 时预测性能更高。这可能是 BP 神经网络模型在处理大样本数据时有优势，用于训练的样本越多，模型训练的越好。

3　讨论

目前，DN 仍然是 21 世纪全球医疗保健的重大临床挑战和负担。一项回顾性研究（含 220 例中国 T2DM 患者）表明，年龄、糖尿病持续时间和 SBP 与 DN 发病风险独立相关[24]，另一项回顾性调查（含 11 771 例 T2DM 患者）显示，较小年龄、高 BMI 和更严重高血压

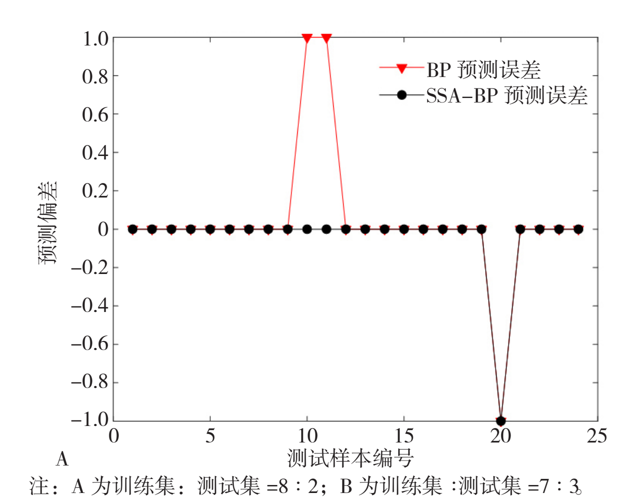


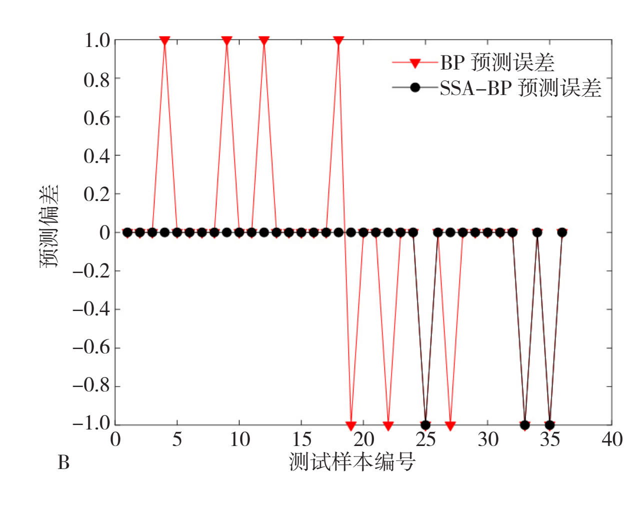
注：A 训练集：测试集 =8：2；B 训练集：测试集 =7：3

图 3　SSA-BP 神经网络模型的进化曲线

Figure 3　Evolutionary curves of SSA-BP neural network model







注：A 为训练集：测试集 =8：2；B 为训练集 :测试集 =7：3。

图 4 BP 神经网络优化前后的预测值和真实值误差对比图

Figure 4 Comparison of predicted and observed value errors before and after BP neural network optimization

表 3 机器学习模型在不同样本拆分比例下预测 DN 的准确率、精确率、灵敏度、特异度、F1-score 和 AUC

Table 3 Accuracy，precision，sensitivity，specificity，F1-score and AUC of machine learning models in predicting DN under varied sample splitting ratios

模型类型			准确率（%）	精确率（%）	灵敏度（%）	特异度（%）	F1-score	AUC
训练集：测试集 =8：2	LR	训练集	89.00	90.00	91.53	85.37	0.907 6	0.884 5
		测试集	83.33	91.67	78.57	90.00	0.846 2	0.842 9
	KNN	训练集	91.00	94.64	89.83	92.68	0.921 7	0.912 6
		测试集	79.17	90.91	71.43	90.00	0.800 0	0.807 1
	SVM	训练集	91.00	94.64	89.83	92.68	0.921 7	0.912 6
		测试集	79.17	90.91	71.43	90.00	0.800 0	0.807 1
	BP 神经网络	训练集	86.00	84.85	93.33	75.00	0.888 9	0.841 7
		测试集	87.50	85.71	92.31	81.82	0.888 9	0.870 6
	SSA-BP 神经网络	训练集	92.00	94.83	91.67	92.50	0.932 2	0.920 8
		测试集	95.83	100.00	92.31	100.00	0.960 0	0.961 5
训练集：测试集 =7：3	LR	训练集	87.50	90.20	88.46	86.11	0.893 2	0.873 0
		测试集	86.11	94.44	80.95	93.33	0.871 8	0.871 0
	KNN	训练集	94.32	97.96	92.31	97.22	0.950 5	0.948 0
		测试集	86.11	94.44	80.95	93.33	0.871 8	0.871 0
	SVM	训练集	89.77	97.78	84.62	97.22	0.907 2	0.909 0
		测试集	86.11	100.00	76.19	100.00	0.864 9	0.881 0
	BP 神经网络	训练集	85.23	92.00	83.64	87.88	0.8762 1	0.857 6
		测试集	72.22	75.00	66.67	77.78	0.705 9	0.722 2
	SSA-BP 神经网络	训练集	94.32	94.64	96.36	90.91	0.955 0	0.936 4
		测试集	91.67	100.00	83.33	100.00	0.909 1	0.916 7

注：LR=Logistic 回归，KNN=K 近邻，SVM= 支持向量机。

是增加 DN 发病率的独立危险因素[25]，这与本研究结果一致。LASSO 回归结果显示，年龄和 DN 发病呈负相关，说明年龄 40~<60 岁者较年龄 <40 岁者更不易患 DN，可能因为 2 型糖尿病在年轻人中更为常见，患有 2 型糖尿病的年轻人表现出典型的一系列危险因素，如不良的生活习惯和环境因素导致的肥胖、胰岛素抵抗、高血压和血脂异常，这些也是 DN 的风险因素[26]。而与本研究结果不同的是，国际糖尿病联盟数据显示糖尿病患病率随年龄增长而增加，年龄范围在 65~79 岁人群的患病率为 19.9%（1.112 亿），达到最高[27]，而 RAVINDRAN 等[28]发现年龄与 DN 之间没有相关性。

高珍秀[29]证实了 HbA_{1c}、SBP 和脉压的变异性是 DN 发生发展的关键影响因子。今日研究小组发现与 DN 等微血管并发症发生风险增加相关的因素是高水平 HbA_{1c}[30]。有研究表明 HbA_{1c} 升高是肾小球滤过率快速下降的危险因素[31]。英国前瞻性糖尿病研究[32]表明长期血糖控制不佳是糖尿病发生微血管并发症或进一步恶化主要危险因素，并且该风险随着 HbA_{1c} 水平的升高

呈指数增加。本研究结果显示 HbA_{1c} 与 DN 正相关，这与前述研究结果一致。既往表明改善血糖控制对 DN 的发生和进展具有有益的作用[33]，然而，SHIKATA 等[34]的研究表明，强化血糖控制对日本 DN 患者并没有显示出治疗优势。强化血糖控制对肾病的益处目前还存在争议[35]。

GALL 等[36]对 26 名患者（1 名女性）平均随访 5.2 年，结果表明 SBP（并非 DBP）升高会加速 T2DM 患者 DN 的进展。SHI 等[37]开展的一项横断面研究（4 219 例患者）结果显示，SBP 是 DN 的危险因素。有研究单因素 Logistic 分析显示总胆固醇（TC）、TG、LDL 对肾功能进展有影响[38]。今日研究小组的研究表明高血压和血脂异常与 DN 发生风险增加相关[30]。本研究 LASSO 回归结果显示 SBP、LDL 与 DN 正相关，LR 模型结果显示 SBP 和 LDL 是 DN 的危险因素，与既往研究一致。

研究表明 DN 随时间的推移而发展，发病高峰出现在患糖尿病 10~20 年后，发病率为 20%~40%[39]。一项来自巴基斯坦随访 12 年的研究表明糖尿病持续时间越长，DN 的发病率越高[40]。JIANG 等[41]以 302 例 T2D 患者为研究对象开发了一个 DN 预测模型，发现典型的 T2DM 患者 DN 的病程通常超过 10 年。SHI 等[37]的研究表明糖尿病病程 >10 年的 T2DM 患者患 DN 的风险较高，其次是病程为 5~10 年的患者。本研究 LR 模型结果显示糖尿病持续时间是 DN 的危险因素，LASSO 回归结果显示，糖尿病持续时间和 DN 正相关，糖尿病持续时间≥ 10 年的患者 DN 的发病风险较高，与既往研究结果一致。

本研究分别采用 LR、KNN、SVM、BP 神经网络、SSA-BP 神经网络建立 DN 诊断模型，总体上 SSA-BP 神经网络模型性能最佳。传统 LR 分析适用范围广，应用灵活。对于特定的问题，其性能相当于甚至优于一些相对复杂的机器学习算法[42]。LYNAM 等[43]在判别糖尿病患者类型（1 型 /2 型）时，LR 模型的性能与更复杂的方法（如神经网络、KNN 模型、随机森林、SVM 模型）一样好。于大海等[44]在评估肝硬化上消化道出血患者的预后时发现 LR 模型的准确率（81.5%）高于决策树（75.1%），本研究之一相同，当训练集：测试集 =8：2 时，在测试集上 LR 模型的准确率（83.33%）高于 KNN 和 SVM，且本研究 LR 模型性能优于于大海等研究，可能原因为本研究样本量虽小但数据代表性强于后者。在对妊娠期糖尿病的早期预测研究中，机器学习模型的整体性能与 LR 模型相似[45]。DAGHISTANI 等[46]基于风险因素预测糖尿病时，随机森林算法的精确率、灵敏度分别为 0.883 和 0.880，预测性能高于 LR 算法（0.692 和 0.703）。本研究 LR 算法精确率和灵敏度分别为 91.67%、78.57%，优于上述研究的 LR 模型。有研究表明，一般对于平衡和不平衡数据，SVM 模型和 LR 模型具有相同的性能，而对于高度不平衡的数据集，SVM 模型可能会更好[47]。但与本研究结果不一致，训练集：测试集 =8：2 时，测试集上 LR 模型的整体预测性能优于 SVM 模型。KNN 被称为惰性算法，因为没有明显的训练阶段，即使有也非常小[48]。训练集：测试集 =8：2 时 KNN 模型的测试集准确率只达到 79.17%，可能是 KNN 模型在训练过程中学习到的东西并不多，效率较低[49]。本研究及既往研究表明，机器学习算法与回归模型的性能结果不一致。分析可能的原因有：LR 模型适用于变量与结果之间具有线性关系的简单数据，而应用于非线性关系的数据性能较差；许多类型的机器学习模型和 LR 模型可能适合不同的数据集，并在不同的数据集中表现不同[45]；同时还有样本量的原因。

近年 BP 神经网络的应用越来越广泛。汪可可等[49]基于 BP 神经网络建立急性脑梗死患者自发性出血性转化的风险预测模型，效果较好。田娟等[50]应用 BP 神经网络构建儿童甲状腺疾病预测模型，结果模型准确度达到 91.43%，误差较小，相比之下本研究 BP 神经网络仅达到了 87.50% 的准确率，但灵敏度较高（92.31%）。然而为克服 BP 神经网络全局搜索能力低下等缺陷，许多研究者对其进行了组合优化研究，并达到了较好的效果。黄仕鑫等[51]使用遗传算法优化 BP 神经网络，建立预测 T2DM 性周围神经病变的模型，准确率分别达到了 98.9%、99.5%，性能优于本研究建立的 SSA-BP 模型。杭昕璇等[52]利用 BP 网络、SSA-BP 网络对麦冬药液糖析出过程建立回归预测模型，发现后者预测精度更高更稳定。韦哲等[53]利用思维进化算法优化的 BP 神经网络判断 2 型糖尿病患者所处的并发症阶段，预测误差更低。本研究中 SSA-BP 神经网络模型在各评估参数上具有明显优势，预测误差更小，优化算法提高了 BP 神经网络的泛化性。

RODRIGUEZ-ROMERO 等[54]预测 2 型糖尿病肾病时（10 251 例）结果显示，RF 模型和 LR 模型性能最好，准确率均为 84.0%，MANIRUZZAMAN 等[12]利用主成分分析进行特征提取，采用线性判别分析、SVM、LR、KNN、朴素贝叶斯和神经网络技术建立 DN 预测模型（133 例），结果高斯核函数（RBF）SVW 分类准确率最高（88.7%）。DAVID 等[55]基于 410 个实例的数据集建立糖尿病肾脏病（DKD）预测模型，结果 KNN 和随机树分类器的性能最好（准确率 93.658 5%）。本研究构建的 SSA-BP 神经网络模型性能优于既往研究，可能因为样本量较小而达到了较好的性能。

神经网络在医学中应用广泛，针对其他糖尿病并发症或慢病，神经网络均可找到适合的网络结构来进行学

习，有较好的拓展性。SSA-BP 神经网络模型无需对输入数据的统计模型作任何先验假设，为基于神经网络的 2 型糖尿病肾病的准确预测提供了算法支持和理论依据。本研究的局限是数据样本量较少，未来可基于大样本数据进行进一步的探索；此外，未进行外部验证，有研究建议模型最好利用外部数据集和领域专家来检查模型的合理性，像支持向量机或神经网络等"黑盒模型"，可解释性差，只能通过外部验证或借助可解释机器学习模型[17]，未来将继续探索其在外部数据集上的性能。

作者贡献：邹琼、张杨进行数据的下载及整理；邹琼、吴曦、陈长生进行文章的构思与设计、论文的修订；邹琼、吴曦、张杨、万毅、陈长生进行研究的实施与可行性分析；邹琼、吴曦、张杨、万毅进行结果的分析与解释并撰写论文。

本文无利益冲突。

参考文献

[1] CHO N H，SHAW J E，KARURANGA S，et al. IDF Diabetes Atlas：global estimates of diabetes prevalence for 2017 and projections for 2045 [J]. Diabetes Res Clin Pract，2018，138：271-281. DOI：10.1016/j.diabres.2018.02.023.

[2] ROSSING P. Prediction，progression and prevention of diabetic nephropathy. The Minkowski Lecture 2005 [J]. Diabetologia，2006，49（1）：11-19. DOI：10.1007/s00125-005-0077-3.

[3] THIPSAWAT S. Early detection of diabetic nephropathy in patient with type 2 diabetes mellitus：a review of the literature [J]. Diab Vasc Dis Res，2021，18（6）：14791641211058856. DOI：10.1177/14791641211058856.

[4] SUN Y L，ZHANG D L. Machine learning techniques for screening and diagnosis of diabetes：a survey [J]. Technical Gazette，2019，26（3）：872-880. DOI：10.17559/TV-20190421122826.

[5] 黄富程，刘德新，曹杰，等. 基于 ABC 优化 BP 神经网络的船舶交通流量预测 [J]. 中国航海，2021，44（2）：78-83.

[6] 李卫华，徐涛，李小梨. 基于人工蜂群的 BP 神经网络算法 [J]. 计算机系统应用，2012，21（5）：195-197，183.

[7] XUE J K，SHEN B. A novel swarm intelligence optimization approach：sparrow search algorithm [J]. Syst Sci Contr Eng，2020，8（1）：22-34. DOI：10.1080/21642583.2019.1708830.

[8] 孙全，孙渊. 基于麻雀搜索算法的 BP 神经网络优化技术 [J]. 上海电机学院学报，2022，25（1）：12-16. DOI：10.3969/j.issn.2095-0020.2022.01.003.

[9] YAN P C，SHANG S H，ZHANG C Y，et al. Research on the processing of coal mine water source data by optimizing BP neural network algorithm with sparrow search algorithm [J]. IEEE Access，2021，9：108718-108730. DOI：10.1109/ACCESS.2021.3102020.

[10] KHODADADI B，MOUSAVI N，MOUSAVI M，et al. Diagnosis and predictive clinical and para-clinical cutoffs for diabetes complications in Lur and Lak populations of Iran：a ROC curve analysis to design a regional guideline [J]. J Nephropharmacol，2018，7（2）：83-89. DOI：10.15171/npj.2018.19.

[11] JIAN Y Z，PASQUIER M，SAGAHYROON A，et al. A machine learning approach to predicting diabetes complications [J]. Healthcare，2021，9（12）：1712. DOI：10.3390/healthcare9121712.

[12] MANIRUZZAMAN M，ISLAM M M，RAHMAN M J，et al. Risk prediction of diabetic nephropathy using machine learning techniques：a pilot study with secondary data [J]. Diabetes Metab Syndr，2021，15（5）：102263. DOI：10.1016/j.dsx.2021.102263.

[13] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. J R Stat Soc，1996，58（1）：267-288. DOI：10.1111/j.2517-6161.1996.tb02080.x.

[14] NIE X，DENG G M. Enterprise financial early warning based on lasso regression screening variables [J]. J Financ Risk Manag，2020，9（4）：454-461. DOI：10.4236/jfrm.2020.94024.

[15] 李阳，陈晓泓，王一梅，等. 基于 LASSO 变量选择联合贝叶斯网络构建恶性肿瘤相关急性肾损伤（AKI）风险预测模型 [J]. 复旦学报：医学版，2020，47（4）：521-530. DOI：10.3969/j.issn.1672-8467.2020.04.009.

[16] MAHESH B. Machine learning algorithms-a review [J]. International Journal of Science and Research（IJSR），2020，9：381-386. DOI：10.21275/ART20203995.

[17] DREISEITL S，OHNO-MACHADO L. Logistic regression and artificial neural network classification models：a methodology review [J]. J Biomed Inform，2002，35（5/6）：352-359. DOI：10.1016/s1532-0464(03)00034-0.

[18] CORTES C，VAPNIK V. Support-vector networks [J]. Mach Lang，1995，20（3）：273-297. DOI：10.1023/A:1022627411411.

[19] JOACHIMS T. Making large-scale SVM learning practical [J]. Technical Reports，1998，8（3）：499-526. DOI：10.1162/153244302760200704.

[20] LARABI-MARIE-SAINTE，ABURAHMAH，ALMOHAINI，et al. Current techniques for diabetes prediction：review and case study [J]. Appl Sci，2019，9（21）：4604. DOI：10.3390/app9214604.

[21] 许条建，金延儒，蒋梅荣，等. 基于麻雀搜索算法优化 BP 神经网络的深远海养殖平台系缆力预报研究 [J]. 渔业现代化，2022，49（6）：17-26. DOI：10.3969/j.issn.1007-9580.2022.06.003.

[22] RAHMAN T，FARZANA S M，KHANOM A Z. Prediction of diabetes induced complications using different machine learning algorithms [D]. Bengal：BRAC University，2018.

[23] LI H S，LAI L，CHEN L，et al. The prediction in computer color matching of dentistry based on GA+BP neural network [J]. Comput Math Methods Med，2015，2015：816719. DOI：10.1155/2015/816719.

[24] WANG X Y，LI J，HUO L X，et al. Clinical characteristics of diabetic nephropathy in patients with type 2 diabetic mellitus manifesting heavy proteinuria：a retrospective analysis of 220 cases [J]. Diabetes Res Clin Pract，2019，157：107874. DOI：10.1016/j.diabres.2019.107874.

[25] MIAO D D，PAN E C，ZHANG Q，et al. Development and validation of a model for predicting diabetic nephropathy in Chinese people [J]. Biomed Environ Sci，2017，30（2）：106-112.



DOI：10.3967/bes2017.014.

[26] HARJUTSALO V，GROOP P H. Epidemiology and risk factors for diabetic kidney disease[J]. Adv Chronic Kidney Dis，2014，21(3)：260–266. DOI：10.1053/j.ackd.2014.03.009.

[27] SAEEDI P，PETERSOHN I，SALPEA P，et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045：results from the International Diabetes Federation Diabetes Atlas，9th edition [J] . Diabetes Res Clin Pract，2019，157：107843. DOI：10.1016/j.diabres.2019.107843.

[28] RAVINDRAN R，KALAIVALLI S，SRINIVASAGALU S，et al. A study on prevalence and risk factors of diabetic nephropathy in newly detected type 2 diabetic patients [J] . J Diabetol，2020，11（2）：109. DOI：10.4103/jod.jod_17_19.

[29] 高珍秀 . 基于深度学习技术的 2 型糖尿病肾病风险预测模型的构建 [D] . 南京：南京中医药大学，2021.

[30] TODAY Study Group，BJORNSTAD P，DREWS K L，et al. Long-term complications in youth-onset type 2 diabetes [J] . N Engl J Med，2021，385（5）：416–426. DOI：10.1056/NEJMoa2100165.

[31] GROUP T S. Effects of metabolic factors，race-ethnicity，and sex on the development of nephropathy in adolescents and young adults with type 2 diabetes：results from the TODAY study [J] . Diabetes Care，2021，45（5）：1056–1064. DOI：10.2337/dc21-1085.

[32] STRATTON I M，ADLER A I，NEIL H A，et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes （UKPDS 35）：prospective observational study [J] . BMJ，2000，321（7258）：405–412. DOI：10.1136/bmj.321.7258.405.

[33] ISMAIL-BEIGI F，CRAVEN T，BANERJI M A，et al. Effect of intensive treatment of hyperglycaemia on microvascular outcomes in type 2 diabetes：an analysis of the ACCORD randomised trial [J] . Lancet，2010，376（9739）：419–430. DOI：10.1016/S0140-6736(10)60576-4.

[34] SHIKATA K，HANEDA M，NINOMIYA T，et al. Randomized trial of an intensified，multifactorial intervention in patients with advanced-stage diabetic kidney disease：diabetic Nephropathy Remission and Regression Team Trial in Japan（DNETT-Japan）[J]. J Diabetes Investig，2021，12（2）：207–216. DOI：10.1111/jdi.13339.

[35] LERMA E V，BATUMAN V. Diabetes and kidney disease M] . Switzerland：Springer，2014：1–679.

[36] GALL M A，NIELSEN F S，SMIDT U M，et al. The course of kidney function in type 2 （non-insulin-dependent） diabetic patients with diabetic nephropathy[J]. Diabetologia，1993，36(10)：1071–1078. DOI：10.1007/BF02374501.

[37] SHI R，NIU Z Y，WU B R，et al. Nomogram for the risk of diabetic nephropathy or diabetic retinopathy among patients with type 2 diabetes mellitus based on questionnaire and biochemical indicators：a cross-sectional study [J] . Diabetes Metab Syndr Obes，2020，13：1215–1229. DOI：10.2147/DMSO.S244061.

[38] 何洋 . 糖尿病肾病进展的危险因素及预测方程的建立 [D] . 兰州：兰州大学，2021.

[39] SAGOO M K，GNUDI L. Diabetic nephropathy：an overview [J] . Methods Mol Biol，2020，2067：3–7. DOI：10.1007/978-1-4939-9841-8_1.

[40] FAWWAD A，MUSTAFA N，ZAFAR A B，et al. Incidence of microvascular complications of type 2 diabetes：a 12 year longitudinal study from Karachi-Pakistan [J] . Pak J Med Sci，2018，34（5）：1058–1063. DOI：10.12669/pjms.345.15224.

[41] JIANG S M，FANG J Y，YU T Y，et al. Novel model predicts diabetic nephropathy in type 2 diabetes [J] . Am J Nephrol，2020，51（2）：130–138. DOI：10.1159/000505145.

[42] 马倩倩，孙东旭，石金铭，等 . 基于支持向量机与 XGboost 的成年人群肿瘤患病风险预测研究 [J] . 中国全科医学，2020，23(12)：1486–1491. DOI：10.12114/j.issn.1007-9572.2020.00.066.

[43] LYNAM A L，DENNIS J M，OWEN K R，et al. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting：application to the discrimination between type 1 and type 2 diabetes in young adults[J]. Diagn Progn Res，2020，4：6. DOI：10.1186/s41512-020-00075-2.

[44] 于大海，李金，罗艳虹，等 . 随机森林模型和决策树模型在肝硬化上消化道出血预后中的应用 [J] . 中国卫生统计，2019，36（2）：162–166.

[45] YE Y Z，XIONG Y，ZHOU Q J，et al. Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data：a retrospective cohort study [J] . J Diabetes Res，2020，2020：4168340. DOI：10.1155/2020/4168340.

[46] DAGHISTANI T，ALSHAMMARI R. Comparison of statistical logistic regression and RandomForest machine learning techniques in predicting diabetes [J] . J Adv Inf Technol，2020：78–83. DOI：10.12720/jait.11.2.78-83.

[47] MUSA A B. Comparative study on classification performance between support vector machine and logistic regression [J] . Int J Mach Learn & Cyber，2013，4（1）：13–24. DOI：10.1007/s13042-012-0068-x.

[48] KHORSHID S F，ABDULAZEEZ A M，MOHSIN A. Breast cancer diagnosis based on k-nearest neighbors：a review [J] . PalArch's Journal of Archaeology of Egypt/Egyptology，2021，18（4）：1927–1951.

[49] 汪可可，武建辉，周莹，等 . 基于 BP 神经网络的急性脑梗死患者自发性出血性转化的风险预测研究 [J] . 中国全科医学，2018，21（12）：1413–1418. DOI：10.3969/j.issn.1007-9572.2017.00.189.

[50] 田娟，朱姝婧，陆强，等 . 基于 BP 神经网络预测儿童甲状腺疾病的模型研究 [J] . 中国医学物理学杂志，2020，37（10）：1340–1344. DOI：10.3969/j.issn.1005-202X.2020.10.022.

[51] 黄仕鑫，浦科学，桑祎莹，等 . 基于 GA-BP 神经网络模型鉴别 2 型糖尿病性周围神经病变的分类模型研究 [J] . 解放军医学杂志，2020，45（1）：73–78. DOI：10.11855/j.issn.0577-7402.2020.01.08.

[52] 杭昕璇，周梓涵 . 基于 SSA-BP 算法的糖析出建模与参数优化研究 [J] . 自动化应用，2022（3）：10–12. DOI：10.19769/j.zdhy.2022.03.004.

[53] 韦哲，石栋栋，王能才，等 . 基于思维进化算法优化的 BP 神经网络对糖尿病并发症的预测研究 [J] . 中国医学装备，

2020，17（10）：1–4. DOI：10.3969/J.ISSN.1672–8270.2020.10.001.
[54] RODRIGUEZ-ROMERO V，BERGSTROM R F，DECKER B S，et al. Prediction of nephropathy in type 2 diabetes：an analysis of the ACCORD trial applying machine learning techniques [J]. Clin Transl Sci，2019，12（5）：519–528. DOI：10.1111/cts.12647.
[55] DAVID S K，RAFIULLAH M，SIDDIQUI K. Comparison of different machine learning techniques to predict diabetic kidney disease [J]. J Healthc Eng，2022，2022：7378307. DOI：10.1155/2022/7378307.

（收稿日期：2023–06–20；修回日期：2023–09–05）

（本文编辑：赵跃翠）